Universität Zürich[UZH]

Institut für Computerlinguistik

# Masterarbeit: Computerlinguistik und Massenüberwachung

Im Lichte der Enthüllungen Snowdens

Hernani Marques Madeira <h2m@access.uzh.ch>
OpenPGP-FP: 7FE5 71F9 3B0C AE18 8424 C7C2 7B83 6E41 F7AB 9CE5
Betreuung: Dr. Noah Bubenhofer (Referent: Prof. Dr. Martin Volk)






# Motivation und Ziele der Masterarbeit

# Motivation (1)

- Persönliches Interesse für Grundrechtsthemen: z. B. Datenschutz und Informationsfreiheit
- Vergegenwärtigung des massiven Überwachungskomplexes: z. B. durch Berichte parlamentarischer Kontrollgremien, Snowden-Enthüllungen und WikiLeaks-Veröffentlichungen
- Politische Entwicklungen hin zu mehr und deutlicheren Formen der Massenüberwachung (auch in der Schweiz: z. B. Gesetze BÜPF und NDG)






# Motivation (2)

- Offenlegung / Entzauberung des «Herrschaftswissens» in Sachen Theorie und Praxis der Massenüberwachung
- Anregung Diskussion Wirksamkeit von Massenüberwachung zur angeblichen Erhöhung der Sicherheit aller (politisches Verkaufsargument)
- Last but not least: Sensibilisierung für Missbrauchspotenzial der Computerlinguistik im Zusammenhang mit Massenüberwachung

# Wichtige Forschungsfragen

- Was ist Überwachung, spezifisch Massenüberwachung?
- Wie kann Massenüberwachung *computerlinguistisch* betrieben werden?
  - Was wird [wurde] erforscht: z. B. im Rahmen der NSA, von Forschungsprogrammen der EU (z. B. INDECT)?
  - Was wird praktiziert: z. B. in der Schweiz (Funk-Massenüberwachung), gemäss Snowden-Enthüllungen oder WikiLeaks-Veröffentlichungen?
- Welche linguistisch verwertbaren Daten fallen bei der Überwachung vollständiger Datenströme überhaupt an?
- Welche konkreten Daten eines Datenstroms sind am interessantesten und reichen aus, um die Natur der Massenüberwachung im Rahmen der Masterarbeit exemplarisch darzustellen?

# Fokus der Arbeit

- Die Arbeit zeigt im Recherche-Teil breit auf, was *computerlinguistisch* möglich ist, auf Basis von Forschung und Überwachungspraxis, die bekannt geworden ist.
- Die Arbeit fokussiert im praktischen Teil rein auf *computerlinguistisch verwertbare* textuellen Inhalte (z. B. PDF-, XML- oder HTML-Material) und bildet einfache (und doch existierende) Möglichkeiten der Volltextüberwachung auf Basis von (kombinierten) Suchbegriffen ab.
- Im ganz besonderen Fokus steht die Filtrierung nach einer bestimmten Überwachungskategorie (z. B. «Gewaltextremismus») der grossen Datenmengen, die anfallen, wenn Funk- oder Kabelverbindungen vollerfasst werden. Weitergehende manuelle oder automatische Auslese entzieht sich dem Umfang der Arbeit.






# Grenzen der Arbeit

- Es werden keine konkreten formalen Suchbegriffe (Meta-Daten: E-Mail-Adressen o. ä.) zur Einengung beigezogen: das real existierende Szenario wird angenommen, rein nach «verdächtiger Sprache» zu suchen, um entsprechend verdächtige Inhalte zu finden.
- Die konkreten / vollständigen / das Zustandekommen der Überwachungsbegriffe (auch: Selektoren) sind eigentliches «Herrschaftswissen» und nicht öffentlich bekannt.
  - Dennoch: aus Enthüllungen und parlamenterischer Kontrollarbeit sind Einzelfälle bekannt geworden.
  - Somit: es werden Annahmen getroffen, wie Selektoren korpusbasiert (statistisch) und zudem linguistisch motiviert zustande kommen müssten.






# Formen von (Massen-)Überwachung

# Formen geheimdienstlicher Überwachung (NDB)

Die Sensoren des NDB

HUMINT

OSINT

COMINT

Partnerdienste

NDB

Verteidigungsattachés

Kantonale Nachrichtendienststellen

IMINT

# Paradigmen der Massenüberwachung

- Fokus auf Meta-Daten: wer mit wem, wann, wie lange usw. kommuniziert
- Fokus auf Inhalte: was konkret (auf Basis von Wortoberflächen) gesagt wird
- Zudem: Kombinationen denkbar, etwa Filtrierung von Nachrichten bestimmter Kommunikationsteilnehmer mit konkretem Inhalt
- Offene Frage: werden Meta-Daten nicht zu eigentlichem Inhalt, wenn damit eine Geschichte erzählt / förmlich konstruiert werden kann?

# Funk-Massenüberwachung nach Schweizer Art: «Onyx»

# Auswertung Schweizer Massenüberwachung: «Achat»

# Speicherfristen Schweizer Massenüberwachung

- Meta-Daten aller Kommunikation muss von *Zugangs*-Providern sechs Monate festgehalten werden: für *rückwirkenden* Zugriff bei Strafverfolgung. (BÜPF Art. 12)
- Bei Onyx-Massenüberwachung können Meta-Daten fünf Jahre gespeichert werden. (VEKF Art. 4 Abs. 3)
- Inhaltsdaten können bei Onyx 18 Monate lang gespeichert bleiben. (VEKF Art. 4 Abs. 2)
- Revision des Gesetzes BÜPF sieht eine Ausweitung der Speicherfrist auf 12 Monate vor; auch Content- und weitere Dienstanbieter müssten *neu* Meta-Daten speichern.
- Neuschaffung eines Nachrichtendienstgesetzes NDG sieht eine Ausweitung der *präventiven* Massenüberwachung auf Glasfaserkabeln vor: angezapft sollen die Daten direkt bei den Providern werden. Auf BÜPF-Daten bestünde neu ebenso Zugriff.



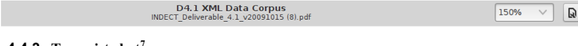
Institut für Computerlinguistik

# Recherche: Forschung zur Massenüberwachung




Institut für Computerlinguistik

# Forschung zur Massenüberwachung: INDECT-Beispiel (1)




### 4.4.3 Terrorist chat[7]

**Shazad Tanweer** [PER.Individual]: Any extra risks getting into **Pakistan** [GPE.NAT] ?

**Omar Khyam** [PER.Individual]: We had five **Bengalis** [GPE.NAT] last year. Guess how **we [PER.Group]** got **them** [GPE.NAT] in. From **Bangladesh** [GPE.NAT] all the way across **India** [GPE.NAT] into **Pakistan**[GPE.NAT]... **we** [PER.Group] bribed the guy [PER.Individual]. You know when you [PER.Individual] go to the check-in, it would all be set up.

**Mohammed Siddique Khan** [PER.Individual]: Going through the airport - normal tickets.

**Omar Khyam**[PER.Individual]: Yeah, just walk straight through bruv normal, just act as if you are a **Pakistani** [GPE.NAT].

**Shazad Tanweer** [PER.Individual]: I live in **Faisalbad** [GPE.NAT]

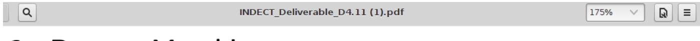
**Omar Khyam** [PER.Individual]: That's not a problem

**Omar Khyam** [PER.Individual]: All right **bruv** [PER.Individual]. Get your parents to pick you up. Or your family ... And that way you will breeze through the airport seriously. Even if **they** [ORG.GOV] are following **you** [PER.Individual] - it doesn't really count. Chill out, proper chill out ... until **we** [PER.Group] contact you and then we'll pick **you** [PER.Individual] up.






# Forschung zur Massenüberwachung: INDECT-Beispiel (2)




## 6. Pattern Matching

In this step we select patterns which show high association to suspicious websites than to normal websites. In many suspicious websites, the sentences containing messages to influence criminal activities are generally grouped within other normal sentences. For example, a suspicious websites can have many factual information and few suspicious lines. Thus, the patterns extracted from such suspicious websites are not all indicative of criminal activities. Most of these patterns will also occur in normal websites. To filter out such normal patterns we use a very simple approach. Once we generate patterns from both suspicious websites and normal websites. The patterns indicative of criminal activities are only those which are not present in normal websites. Thus, we select only patterns which are present in suspicious websites but not in normal websites. For exam-

| Patterns from suspicious websites | Patterns from normal website |
|---|---|
| hand-package-boss | everest-mountain |
| everest-mountain | tall-mountain-world |
| tall-mountain-world | temperature-cold-winter |

Table 4: Possible patterns generated from suspicious and normal websites

# Forschung zur Massenüberwachung: Beispiel NSA-Patente mit CL-Bezug (1)

- *Method of retrieving documents that concern the same topic* (1995)
- *Language-independent method of generating index terms* (1998)
- *Automatically generating a topic description for text and searching and sorting text by topic using the same* (1999)
- *Device and method for full-text large-dictionary string matching using n-gram hashing* (2001)
- *Method for finding large numbers of keywords in continuous text streams* (2001)
- *Method of summarizing text using just the text* (2005)

# Forschung zur Massenüberwachung: Beispiel NSA-Patente mit CL-Bezug (2)

- *Method of optical character recognition using feature recognition and baseline estimation* (2008)
- *Natural language database searching using morphological query term expansion* (2010)
- *Method of database searching* (2010)
- *Method of identifying topic of text using nouns* (2010)
- *Method of assessing language translation and interpretation* (2012)
- *Device for and method of language processing* (2013)






# Recherche: Praxis von Massenüberwachung






# Praxis von Massenüberwachung: XKeyscore-Beispiel (1)

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

## Communication Based Contexts

| | |
|---|---|
| email_body(expr) | The UTF-8 normalized text of all email bodies.<br>email_body('how to' and 'build' and ('bomb' or 'weapon')) |
| chat_body(expr) | The UTF-8 normalized text of all chat bodies.<br>chat_body('how to' and 'build' and ('bomb' or 'weapon')) |
| document_body(expr) | The UTF-8 normalized text of the Office document. – Office documents include (but are not limited to) Microsoft Office, Open Office, Google Docs and Spreadsheets.<br>document_body('how to' and 'build' and ('bomb' or 'weapon')) |
| calendar_body(expr) | The UTF-8 normalized text of all calendars. An example is Google Calendar.<br>calendar_body('wedding') |
| archive_files(expr) | Matches a list of files from within an archive. For example is a ZIP file is transmitted, all names of files within are passed to this context.<br>archive_files('bad.dll' or 'virus.doc') |
| http_post_body(expr) | The UTF-8 normalized text HTTP url-encoded POSTs.<br>http_post_body('action=send' and 'badguy@yahoo') |

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

# Praxis von Massenüberwachung: XKeyscore-Beispiel (2)

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

## Example 4

- $acwitems = 'machine gun' or 'grenade' or 'AK 47'
- $acwpositions = 'minister of defence' or 'defense minister'
- $acwcountries = 'somalia' or 'liberia' or 'sudan'
- $acwbrokers = 'south africa' or 'serbia' or 'bulgaria'
- $acwports = 'rangood' or 'albasra' or 'dar es salam'

```
topic('wmd/acw/govtorgs') =
    email_body($acwitems and $acwpositions and
  ($acwcountries or $acwbrokers or $acwports));
```

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

# Praxis von Massenüberwachung: XKeyscore-Beispiel (3)

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

## Finding Targets




- How do I find a strong-selector for a known target?
- How do I find a cell of terrorists that has no connection to known strong-selectors?
- Answer: Look for anomalous events
  - E.g. Someone whose language is out of place for the region they are in
  - Someone who is using encryption
  - Someone searching the web for suspicious stuff

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

# Praxis von Massenüberwachung: NSA-Untersuchungsausschuss (1)

https://netzpolitik.org/2015/live-blog-aus-dem-geheimdienst-untersuchungsausschuss-2/

k access, place your bookmarks here on the bookmarks bar. Import bookmarks now...

**Sensburg:** Warum heißt das Wortbank-Gruppe, weil Datenbank von Worten, die interessant sind?

**K.M.:** Ja, quasi. Wir hatten schonmal „Bomb" als Suchbegriff drin, das war großer Reinfall, weil auch „Sexbomb" getroffen hatte. Was an arabischen und kyrillischen Sachen in Datenbank ist, ist auch vernachlässigbar.

**Sensburg:** Zurück zu dritter Ebene? Gibt es eine Erklärung zu Begriffen, die ihn als G-10-Hinweis einordnen lassen? Was sind deutsche und europäische Interessen?

**K.M.:** Z.B. Firmennamen.

**Sensburg:** Wie erkennt man europäische Suchbegriffe?

# Praxis von Massenüberwachung: NSA-Untersuchungsausschuss (2)

https://netzpolitik.org/2015/live-blog-aus-dem-geheimdienst-untersuchungsausschuss-2/

k access, place your bookmarks here on the bookmarks bar. Import bookmarks now...

**Sensburg:** Versuche zu verstehen, mit welcher Sensibilität man an die Wahrung deutscher Interessen herangegangen ist. 30.000 erscheint mir wenig, wenn nicht gleich alle Oettingers der Welt rausgefiltert werden.

**K.M.:** Angenommen, NSA steuert nur ganze Mailadressen zu. Wir würden dann mit „siemens.com" Tausende Adressen erschlagen.

**Sensburg:** Warum nicht nur „Siemens"?

**K.M.:** Bei Siemens würde es Sinn machen, bei „Audi" nicht, sonst ist das ja auch „Saudi"



Institut für Computerlinguistik

# Eigene (exemplarische) Massenüberwachung

# Überwachungssetting (1)

- Trainingsdaten: allen Text (auch z. B. aus PDFs) von aufbau.org und pnos.ch (getrennt); Grund: Gruppen gelten als namentlich «gewaltextremistisch» gemäss Bericht vom Nachrichtendienst des Bundes NDB.
- Evaluationsdaten:
  1. Vollüberwachung des eigenen Internetverkehrs für mindestens zehn Tage.
     - Technik: Alix-Board (AMD Geode) als bridge mit FreeBSD und tcpflow
     - Idee: Aufzeigen, ob Treffer erfolgen, am Beispiel des Anschlusses einer “unbescholtenen” Kleinfamilie; und inwiefern diese «extremistisch» (links oder rechts) erfolgen.

# Überwachungssetting (2)

- Evaluationsdaten:
  2. Nutzung öffentlich nutzbarer Suchmaschinen zur Evaluation der Überwachungsbegriffe nach Inhalten, die True (TP) und False Positive (FP) im Sinne verdächtigen Materials nach Trainingskorpus sein können.
  3. Hiermit: Simulation von XKeyscore mit öffentlich indexiertem Material und und aufzeigen, welche Art von Inhalten / Webseiten angesteuert werden müssten, um bei der Massenüberwachung in die engere Auswahl zu geraten.
     - Whitenet: Google, Yahoo!, Bing, MetaGer, Startpage, DuckDuckGo, Swisscows, search.ch
     - Whitenet (peer-to-peer): YaCy
     - Darknet: Not Evil (im Tor-Netzwerk: `hss3uro2hsxfogfq.onion`)

# Überwachungssetting (3)

- Trainings- und eigens erstelltes Evaluationsmaterial (aus überwachtem Datenstrom) wird mittels Apache Tika nach Inhalt gescrapped.
- Doppelte Dateien werden gelöscht.
- Inhaltlich fortbestehendes Navigations-/Fussnoten-/Titelmaterial oder andere nicht-linguistische Artefakte werden auf Basis doppelten Vorkommens (im Vergleich) entfernt.
- Beim Trainingsmaterial wird zudem mittels Apache Tika sichergestellt, dass nur (mehrheitlich) deutsche Dokumente beibehalten werden.
- Evaluation wird mittels Kommandozeilentools, kleinen Skripten und/oder Webbrowser durchgeführt.

# Drei Filtermodelle zur Massenüberwachung (1)

**TF*IDF-Modell:** Überwachung auf Basis relativ häufiger und gut über die Dokumente verteilter, aber *nicht* häufigster und überall (in allen Dokumenten) vorkommender Wörter; ohne Stoppwörter

- Auswahl von **15 Selektoren**; zu je fünf:
  - Einzelworte
  - Wortkombinationen
  - Wort-2-Gramme

# Drei Filtermodelle zur Massenüberwachung (2)

**Verdachtssprache-Modell**: Überwachung auf Basis verdächtiger Sprache mit Fokus auf Skandalvokabular, Sprachintensivierung und -relativierung:

- Paper: Sarah Ebling / Joachim Scharloth / Tobias Dussa / Noah Bubenhofer (2012): *Gibt es eine Sprache des politischen Extremismus?* In: Frank Liedtke (Hrsg.): Sprache, Politik, Partizipation. Bremen: Hempen.
- Auswahl von **15 Selektoren**; zu je fünf:
  - Häufige Einzelworte oder Phrasen mit skandalisierendem Charakter
  - Häufige Relativierer-Wort-Kombinationen
  - Häufige Intensivierer-Wort-Kombinationen

# Drei Filtermodelle zur Massenüberwachung (3)

**LDA-Modell**: Überwachung auf Grund thematischer Ähnlichkeit von Dokumenten; Thema als Kombination verschiedener Wörter einer Textkollektion.

- Auswahl von **25 Selektoren**: Bestimmung von fünf Topics und Auswahl der jeweils fünf wahrscheinlichsten Wörter, um *Kombinationen* von 1–5 Wörter zu bilden.

# Evaluationsarbeit

– In der Summe der drei Modelle zur Massenüberwachung werden 55 Selektoren je Trainingskorpus (aufbau.org; pnos.ch) evaluiert: **insgesamt 110 Selektoren**.
– Es werden zehn Suchmaschinen als Evaluationskorpora genutzt und der eigens überwachte Datenstrom als weiteren (persönlich differenzierten) Evaluationskorpus beigezogen: **insgesamt elf Evaluationskorpora**.
– Im Produkt ergeben sich 1'210 zu vollziehende Selektionen.
  – Bei den Suchmaschinen werden je Selektion fünf Ergebnisse manuell ausgewertet und das Verhältnis TP–FP gemessen.
  – Beim überwachten Datenstrom wird analog vorgegangen, wenn auch weniger Treffer möglich sind (wesentlich kleineres Korpus).
  – Insgesamt werden – gegeben Treffer – **bis zu 6'050** Ergebnisse als verdächtige Dokumente *manuell* evaluiert.

# Einzelbeispiele von möglichen Selektoren

- aufbau.org
  - aufbau.org (TF*IDF-Modell): `repression`, `frauenkampf`, `bulle`
  - aufbau.org (Skandalvokabular): `steuergeschenk`, `hetzerisch-rassistisch`
  - aufbau.org (Intensivierung): `protest (lautstark)`
  - aufbau.org (Relativierung): `demokratie (deckmantel von)`
- pnos.ch
  - pnos.ch (TF*IDF-Modell): `ahnensturm`, `sicherheitsdienst`
  - pnos.ch (Skandalvokabular): `maulkorbgesetz`, `flüchtlingsflut`
  - pnos.ch (Intensivierung): `masseneinwanderung (katastrophal)`
  - pnos.ch (Relativierung): `künstler (sogenannt)`






# Fragen & Kritik zum Beitrag






# Fragen &Kritik